Development of gene signatures: a reality check

Willem Talloen

Outline

- Gene signatures
  - What, the impact & boom
- Why combining genes in a signature
- Analogy with morphology
- Used algorithms
  - Complex solution for a simple problem
  - Most often inappropriate solution
- Simulation illustration
- Conclusion

Gene signatures

A condition's gene signature is the group of genes in a type of cell whose combined expression pattern is uniquely characteristic of that condition.

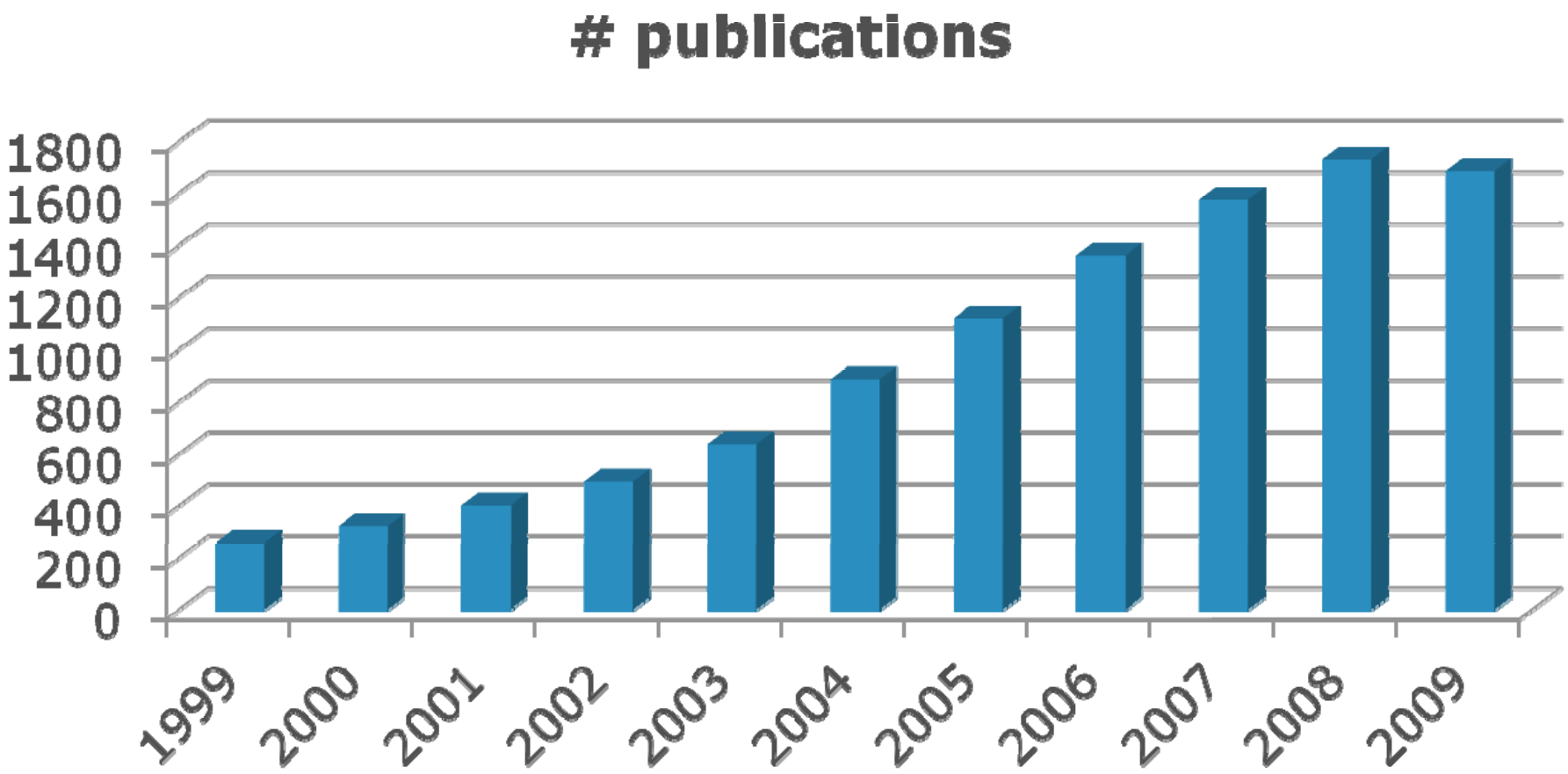
- Examples of conditions
  - Response to therapy
  - Disease
  - Prognosis
- The gene signatures are captured by devices called microarrays that take a snapshot of the tens of thousands of genes at the heart of every cell.

Gene signatures

- One of the most ambitious spinoffs of the human genome project is a new, systematic approach to drug discovery that matches diseases with potential treatments using a universal language based on cells' distinctive gene activity profiles, or "signatures."*
- Big impact in field of 'Personalized medicine'
- Many publications

*Medical News Today 2006

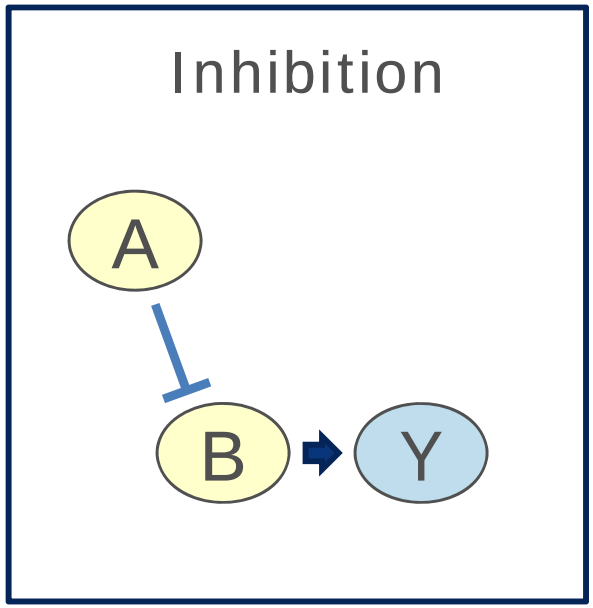
“Gene signature” in pubmed abstracts

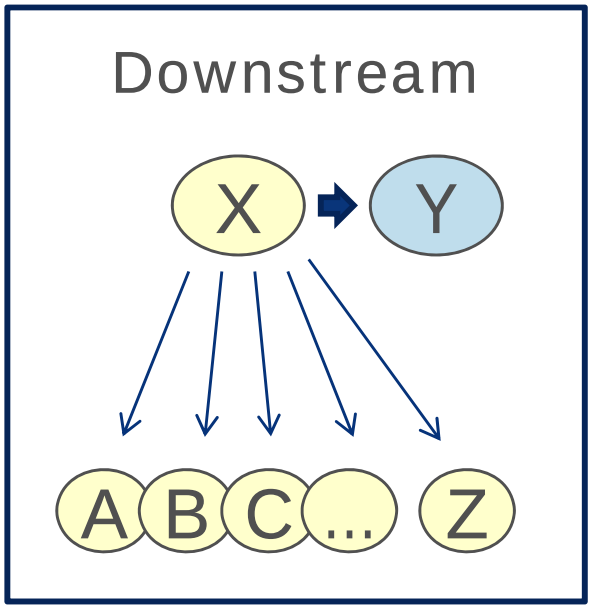


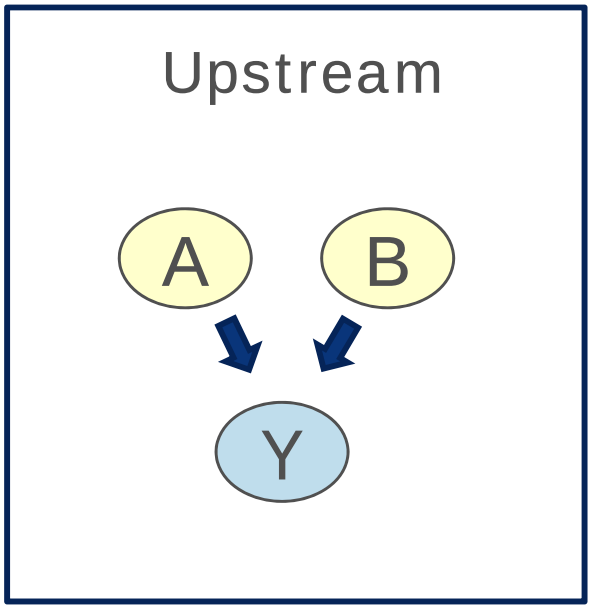
When do multiple markers outperform single markers?

1. Inhibition/catalyzation
   - The genes interact in such a way that their relative proportion is marking the endpoint of interest.
2. Downstream effects
   - Borrowing strength of markers within a pathway. The genes are coregulated and belong to the same pathway that marks the endpoint of interest. Combining the expression levels of multiple genes improves the robustness and the predictive accuracy of the biomarker.
3. Upstream effects
   - There may be multiple causes for a sample to show the phenotype. These different upstream causes should be integrated into a signature to make it more general.

Why a multiple marker signature?







$$\frac{B}{A}$$

$$mean(A, B, ..., Z)$$

$$either\ A\ or\ B$$

Specific examples

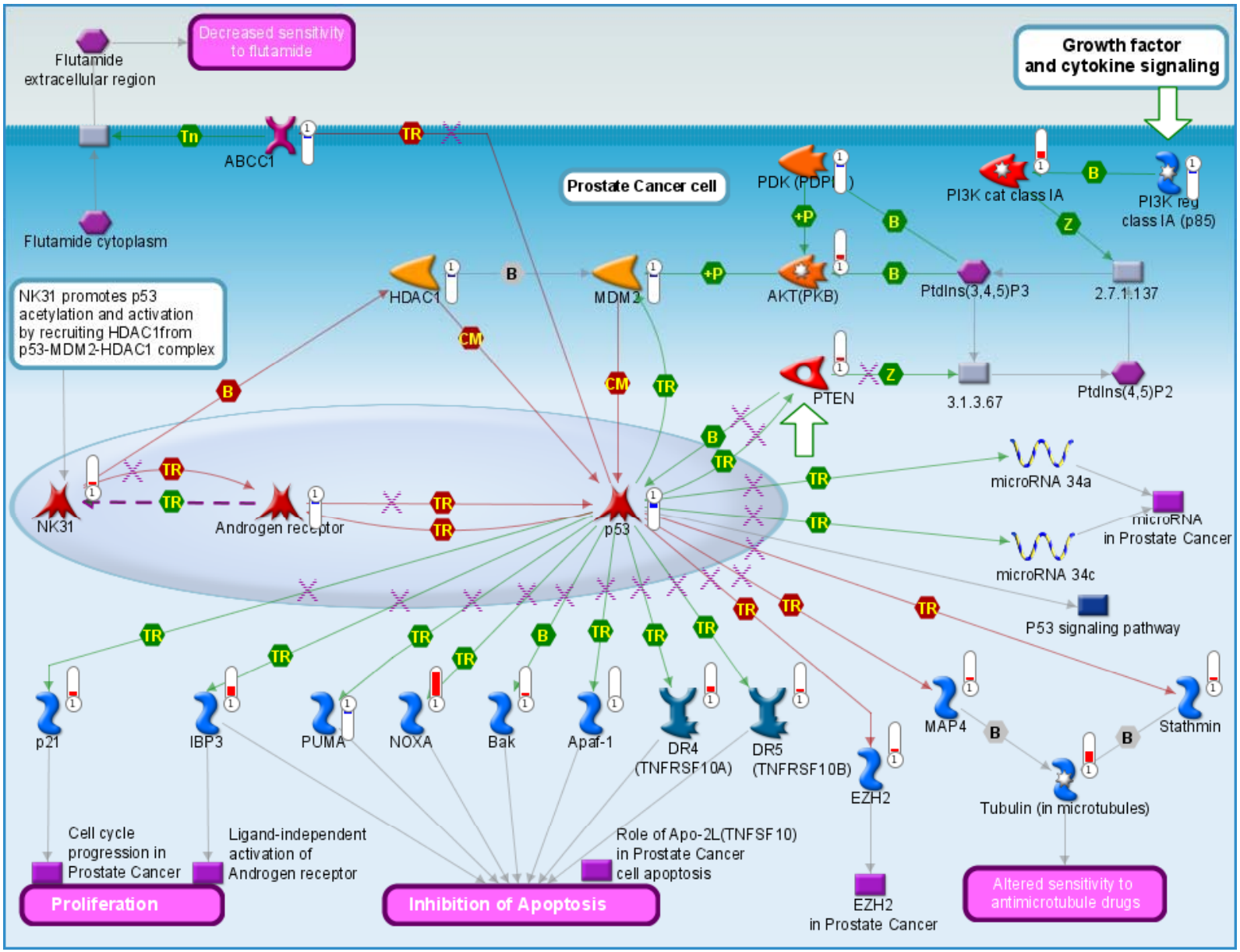
Downstream

- P53 signalling
- KRAS signalling
- IL6 signalling
- …

Inhibition

- Protein/Creatinine Ratio (PCR).
  - The 2005 UK Chronic Kidney Disease guidelines states that PCR is a better test than 24 hour urinary protein measurement.
- …

p53


Flutamide extracellular region
Decreased sensitivity to flutamide
Growth factor and cytokine signaling
Tn
ABCC1
TR
Prostate Cancer cell
PDK (PDPK1)
PI3K cat class IA
B
PI3K reg class IA (p85)
Z
+P
Flutamide cytoplasm
HDAC1
MDM2
AKT(PKB)
PtdIns(3,4,5)P3
2.7.1.137
NK31 promotes p53 acetylation and activation by recruiting HDAC1from p53-MDM2-HDAC1 complex
CM
PTEN
3.1.3.67
PtdIns(4,5)P2
NK31
Androgen receptor
p53
microRNA 34a
microRNA in Prostate Cancer
microRNA 34c
P53 signaling pathway
p21
IBP3
PUMA
NOXA
Bak
Apaf-1
DR4 (TNFRSF10A)
DR5 (TNFRSF10B)
EZH2
MAP4
Stathmin
Tubulin (in microtubules)
Cell cycle progression in Prostate Cancer
Ligand-independent activation of Androgen receptor
Role of Apo-2L(TNFSF10) in Prostate Cancer cell apoptosis
EZH2 in Prostate Cancer
Proliferation
Inhibition of Apoptosis
Altered sensitivity to antimicrotubule drugs

Specific examples

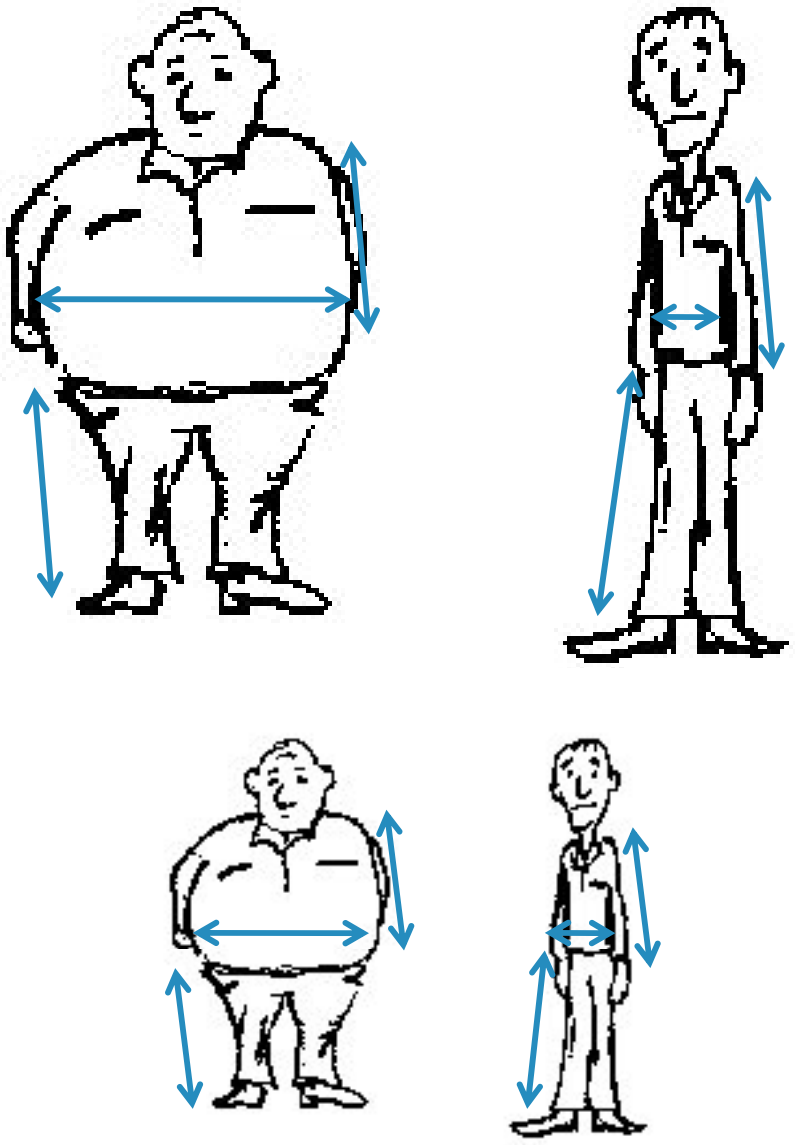
Downstream

- P53 signalling
- KRAS signalling
- IL6 signalling
- …

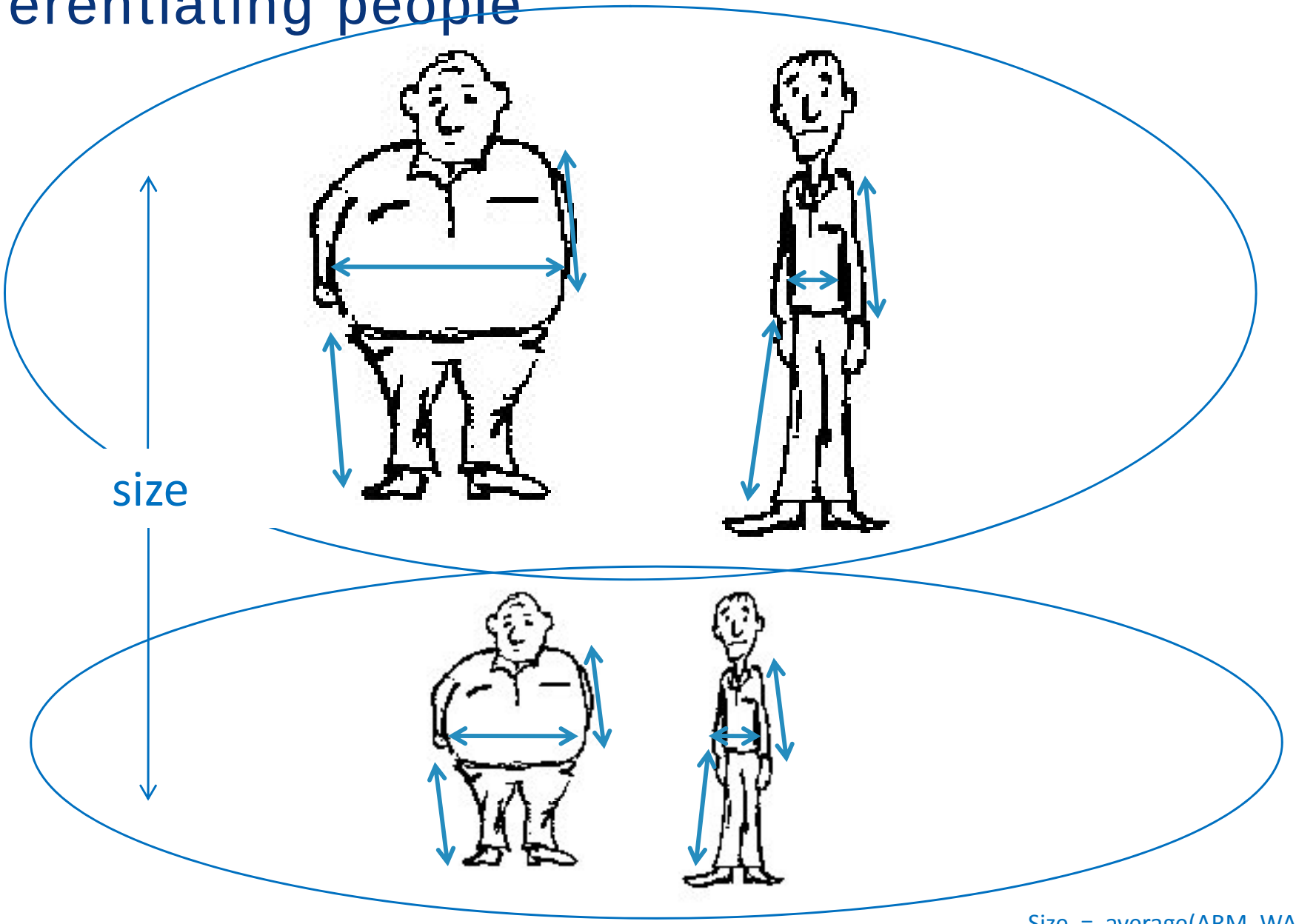
Inhibition

- Protein/Creatinine Ratio (PCR).
  - The 2005 UK Chronic Kidney Disease guidelines states that PCR is a better test than 24 hour urinary protein measurement.
- …

Analogy with Morphology: Differentiating people

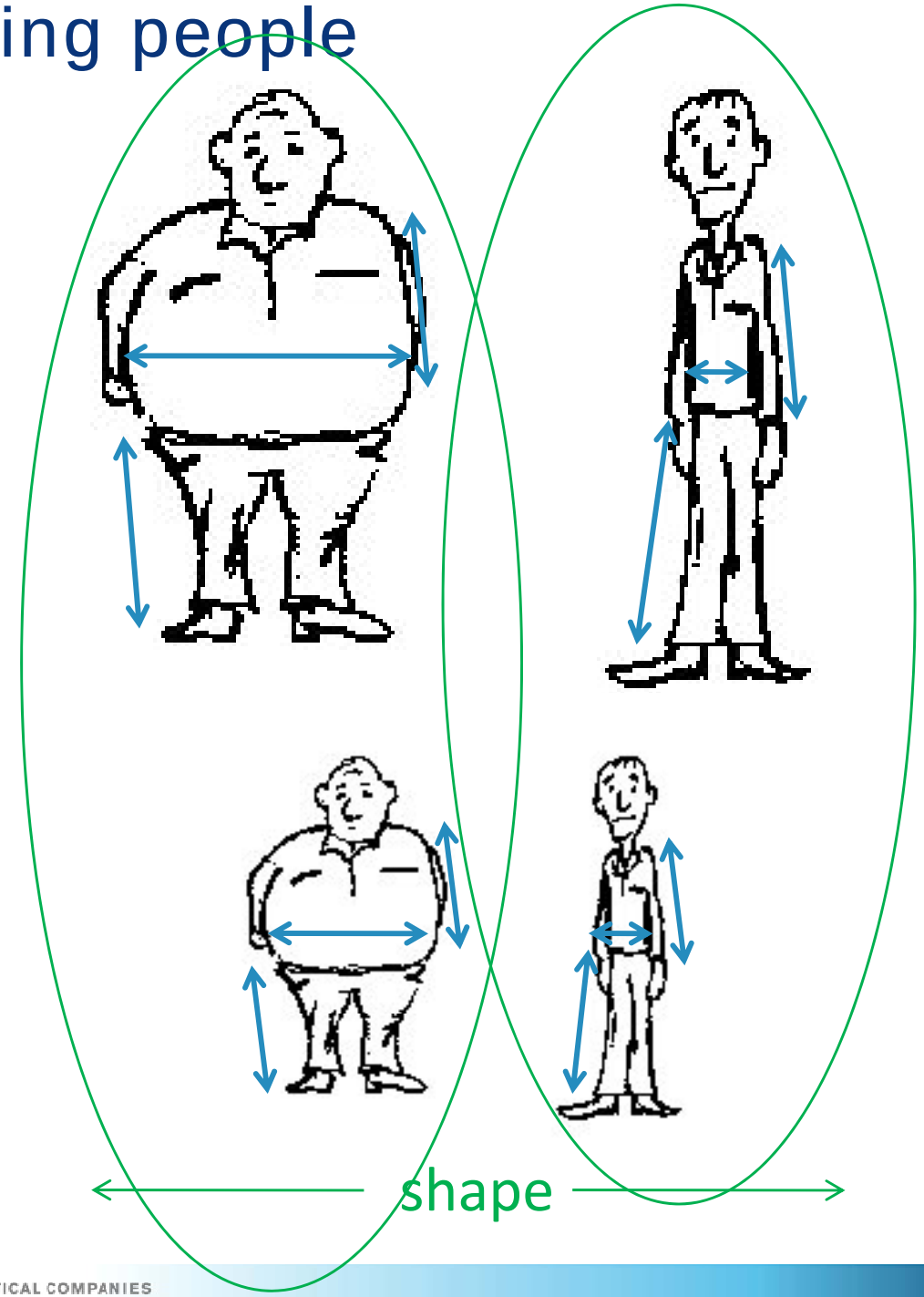


Differentiating people



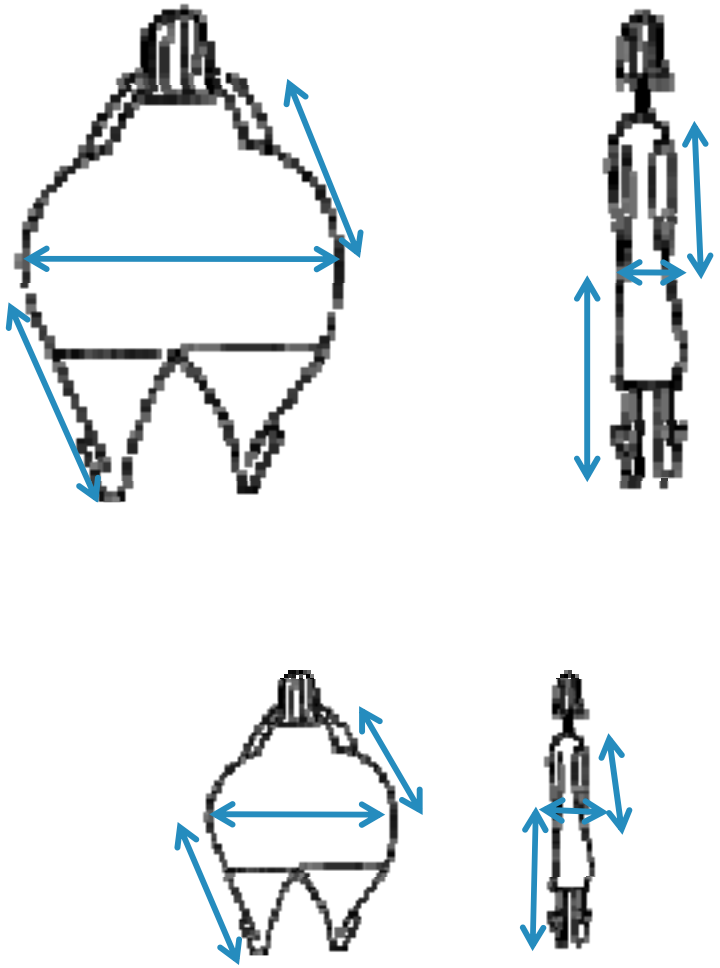
Size = average(ARM, WAIST)

Differentiating people

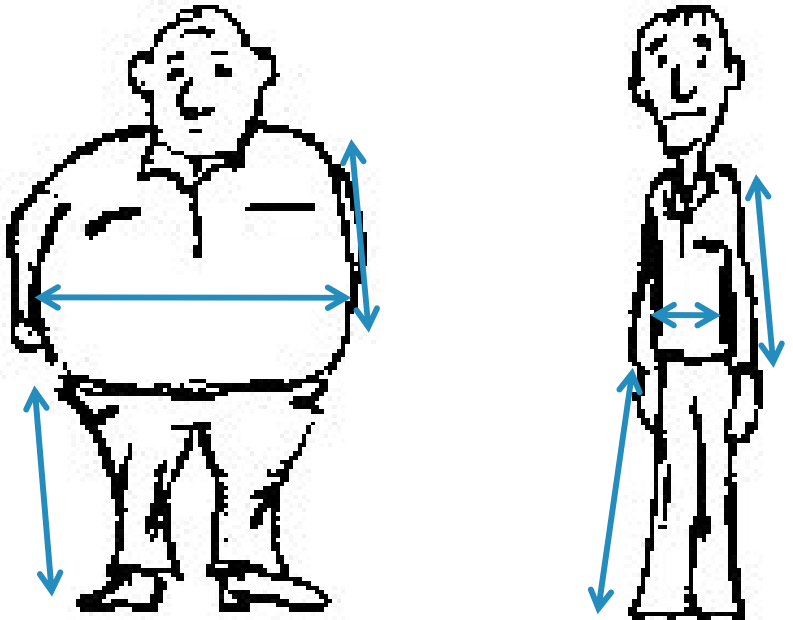


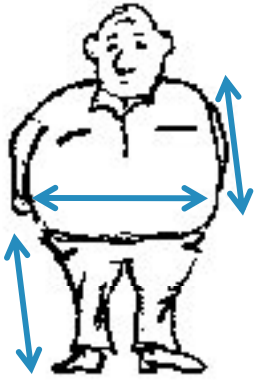
Shape = ARM / WAIST

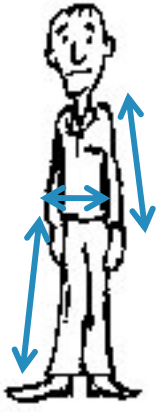
Differentiating people



Differentiating people







Size = average(ARM, WAIST)

size

Downstream effects

Inhibition/ catalysation

shape

Shape = ARM / WAIST

How to combine morphological traits in an index

- Linear combinations after log transformation

$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$

$$Signature = \beta_1 * arm + \beta_2 * leg + \beta_3 * waist$$

– Size: $all\ \beta_i > 0$ — 1st PC of PCA

Weighted average

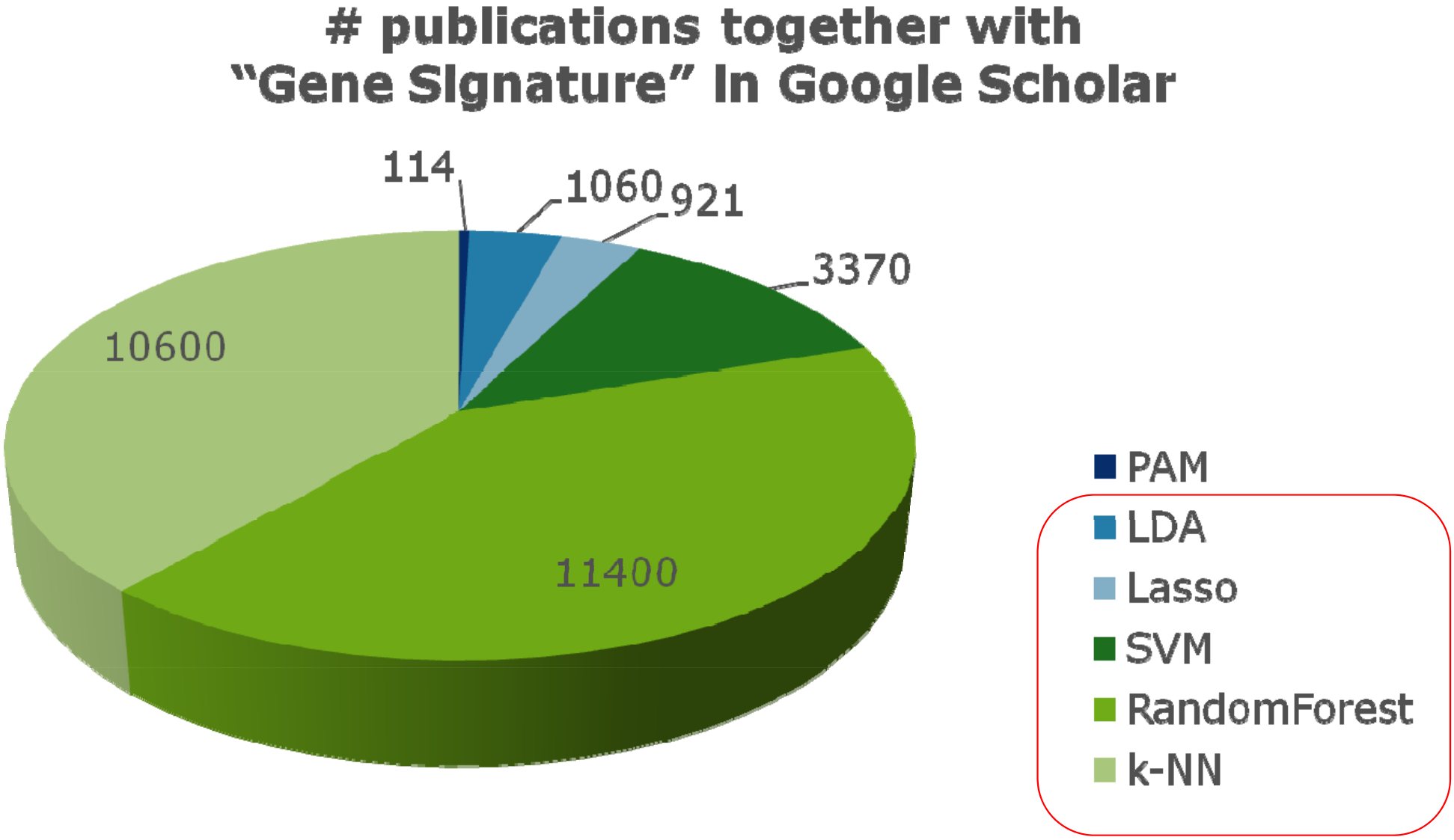
– Shape: $\begin{cases} \beta_1 < 0\ and\ \beta_2 < 0 \\ \beta_3 > 0 \end{cases}$ — LDA

Additive effects

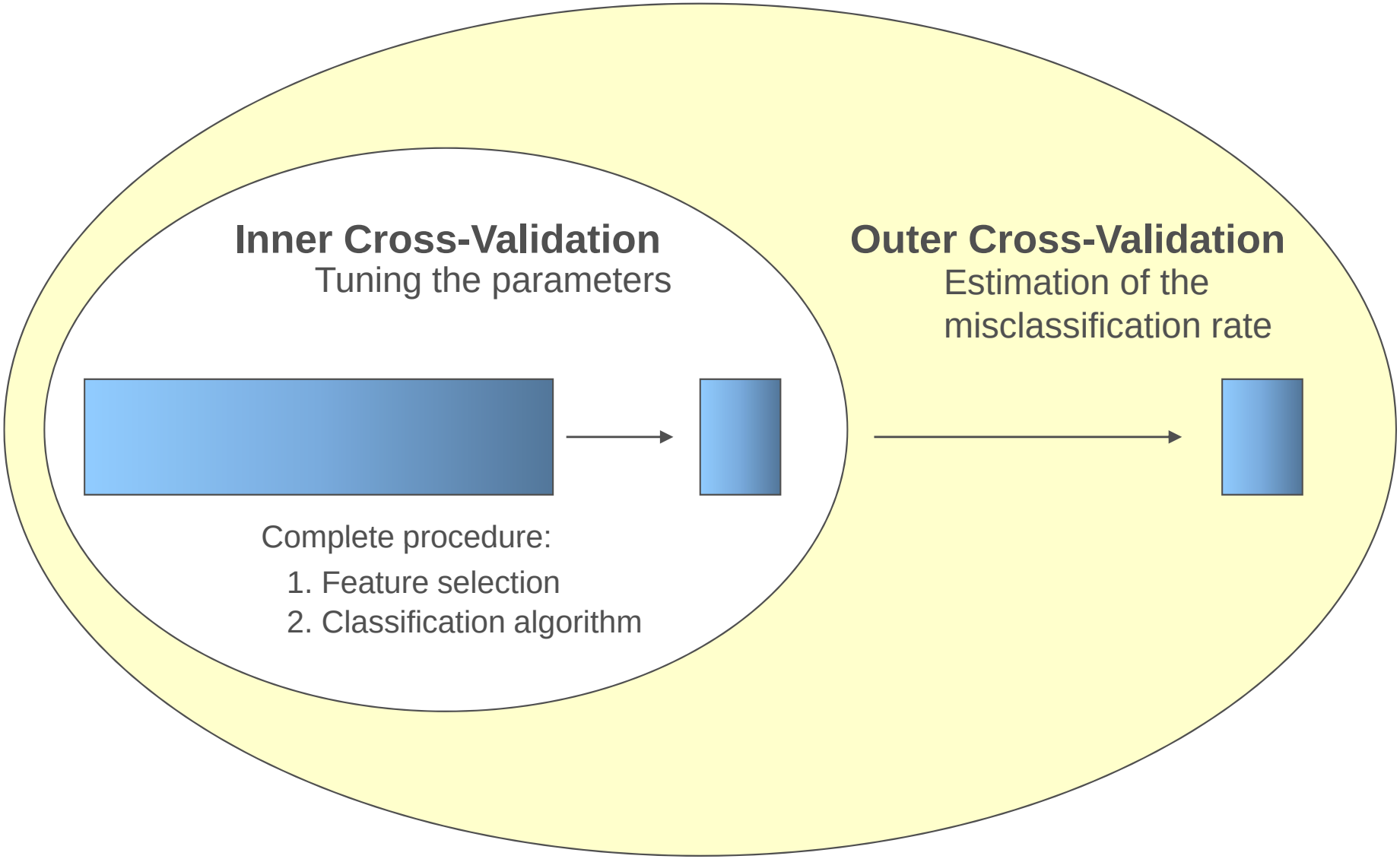
Signature data analysis

1. Feature selection
   - T-test
   - Wilcoxon
   - …
2. Classification
   - Prediction Analysis of Microarrays
   - Discriminant analysis
   - L_1 regularized regression (Lasso)
   - Random Forest
   - Support Vector Machines
   - …

Used algorithms



Cross validation



Hypothesis-driven classification

- Downstream
  1. Feature selection: gene by gene analysis
  2. Composite index (average) of top genes
- Inhibition
  - Linear model
- Upstream
  - Classification tree

Why rarely hypothesis-driven?

- Biological hypothesis formulation rare in Omics experiments
  - Exploratory searches
  - Pathway knowledge is far from comprehensive
- Omics data properties imposed new and interesting statistical challenges.

→This enthusiasm made many researchers forget to think about the biological relevance of these developed classification algorithms.

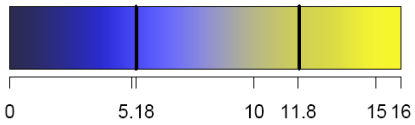
"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

John Tukey (†2000)

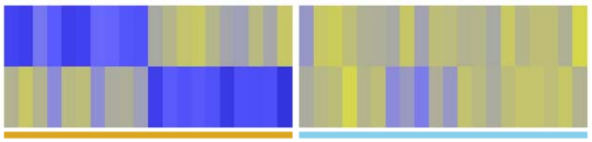
Simulation illustration

- Random data
  - 40 samples (2 groups x 20)
  - 1000 genes
- Downstream:
  - 50 differentiating genes
- Inhibition:
  - 2 genes which ratio differentiates
- Upstream:
  - Samples are differentiated either by gene 1 or by gene 2

Simulation

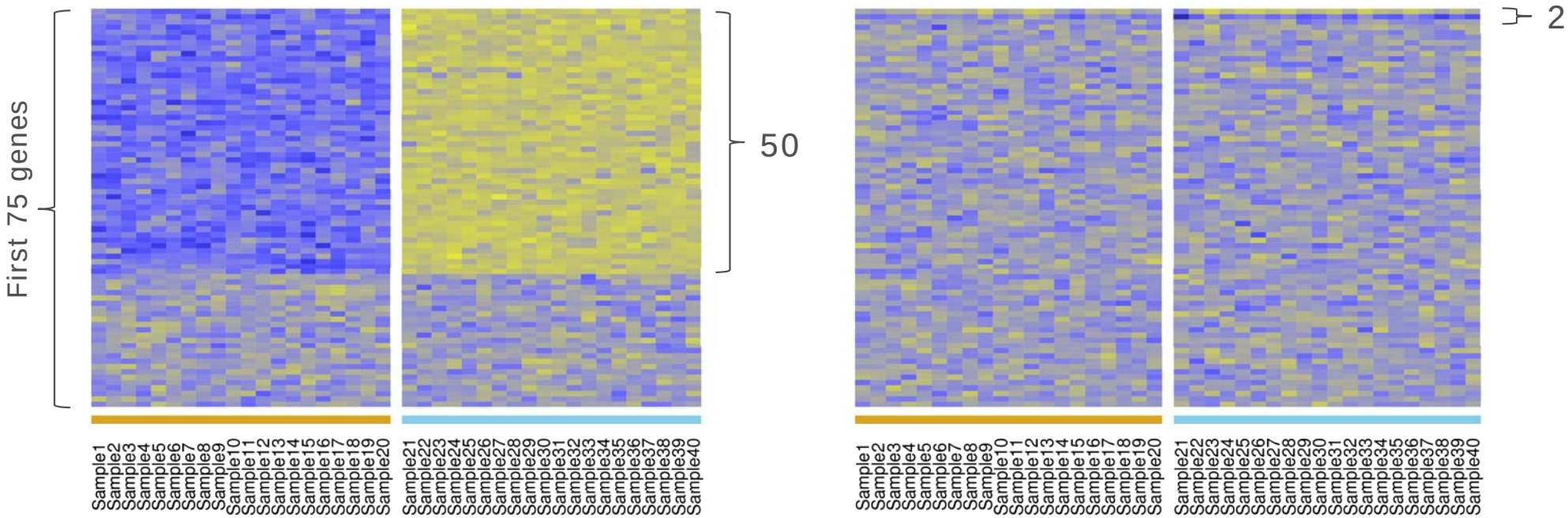


Upstream



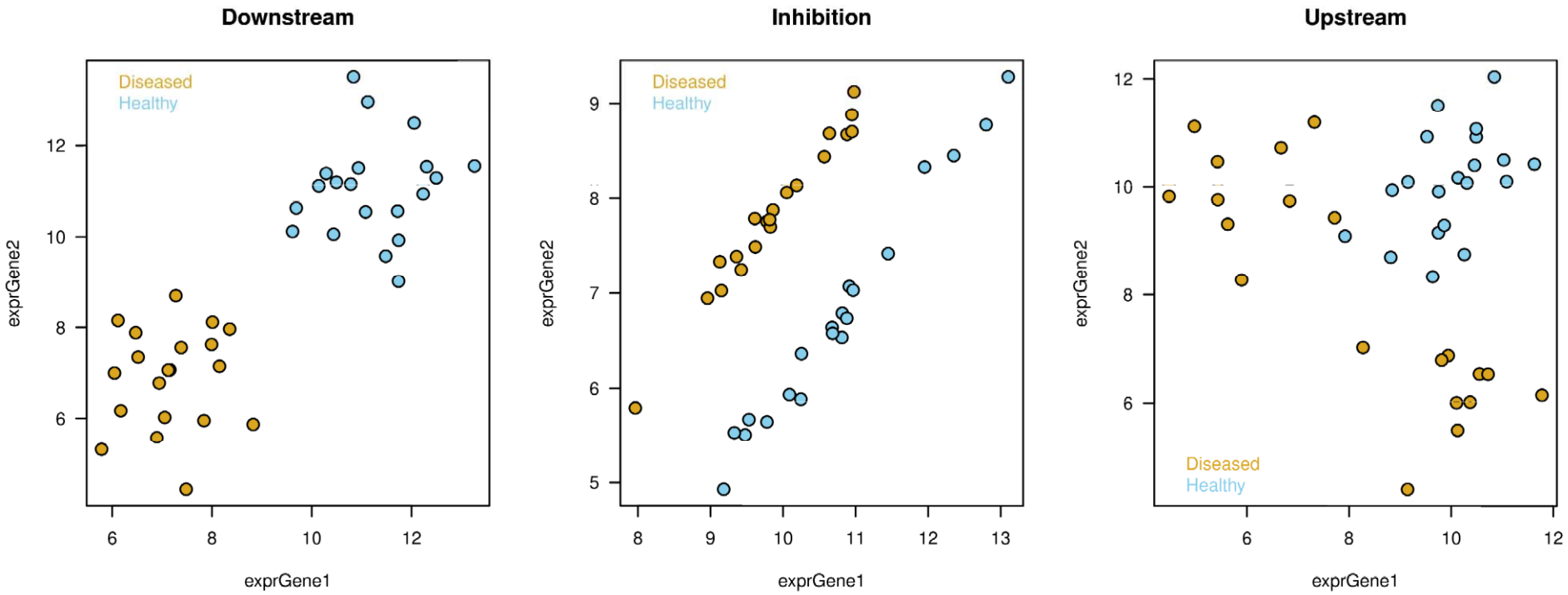
Downstream

Inhibition

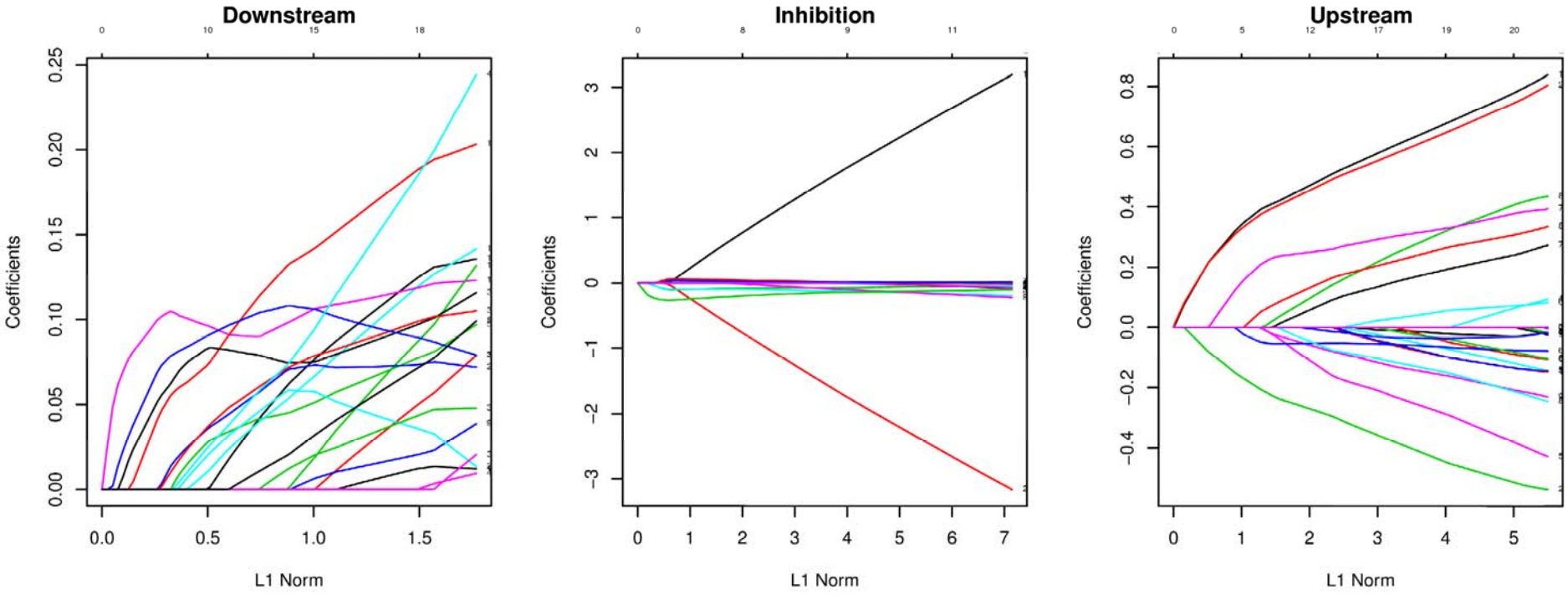


Diseased Healthy Diseased Healthy

Plot of top two genes



Lasso

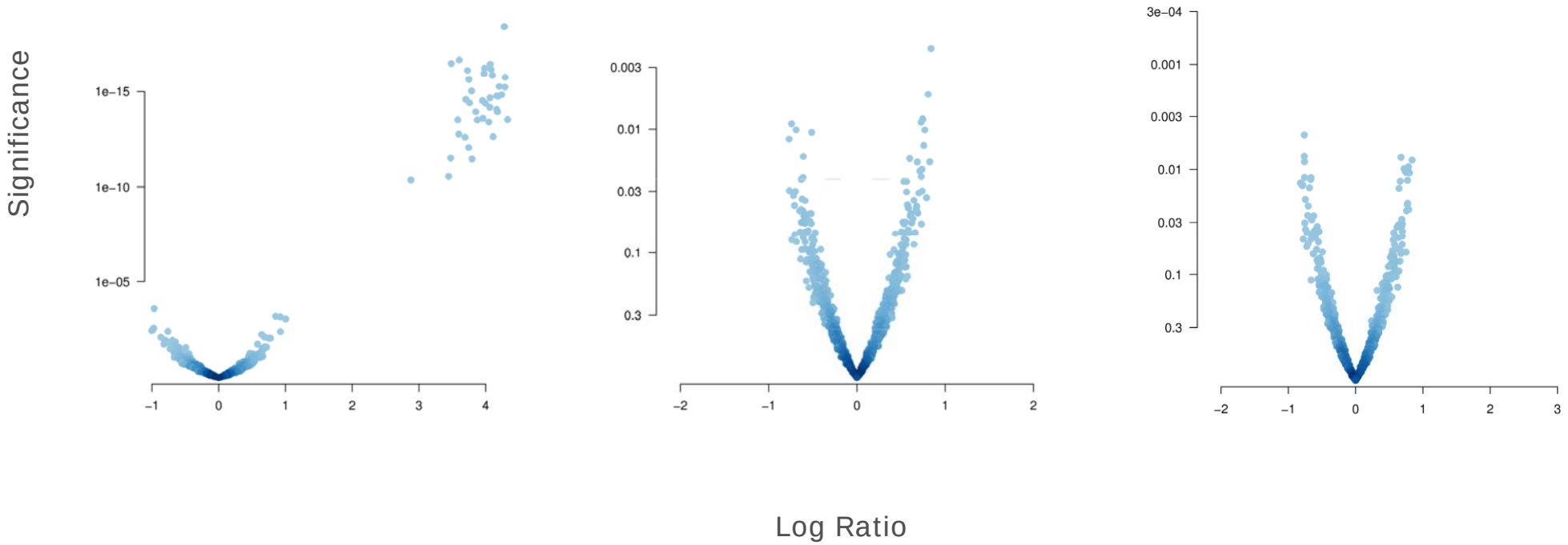

Downstream
Coefficients
L1 Norm
Inhibition
Coefficients
L1 Norm
Upstream
Coefficients
L1 Norm

gene by gene t-tests

Downstream

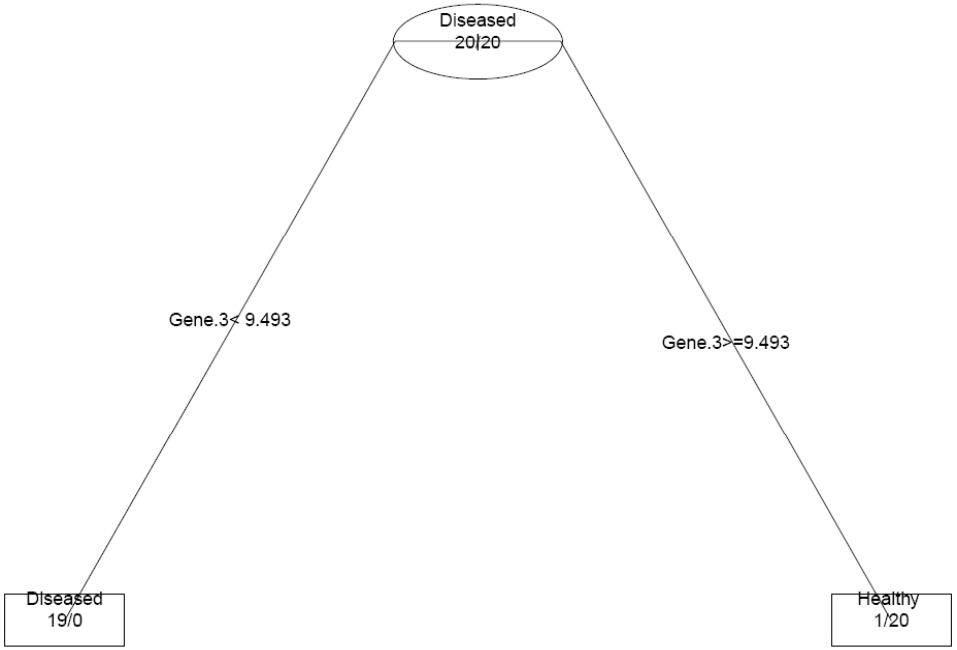
Inhibition

Upstream

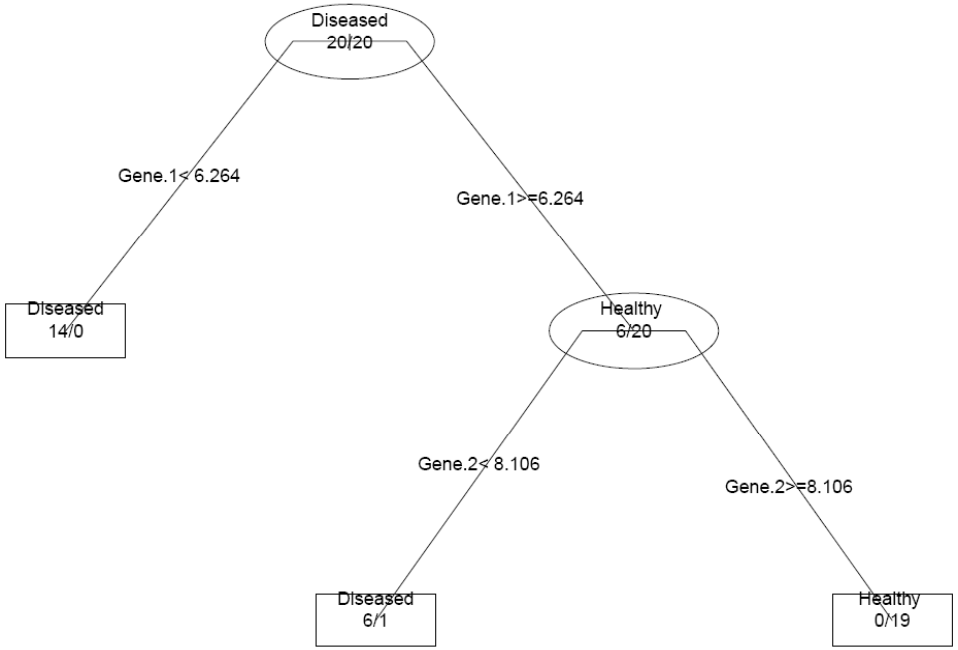


Recursive partioning

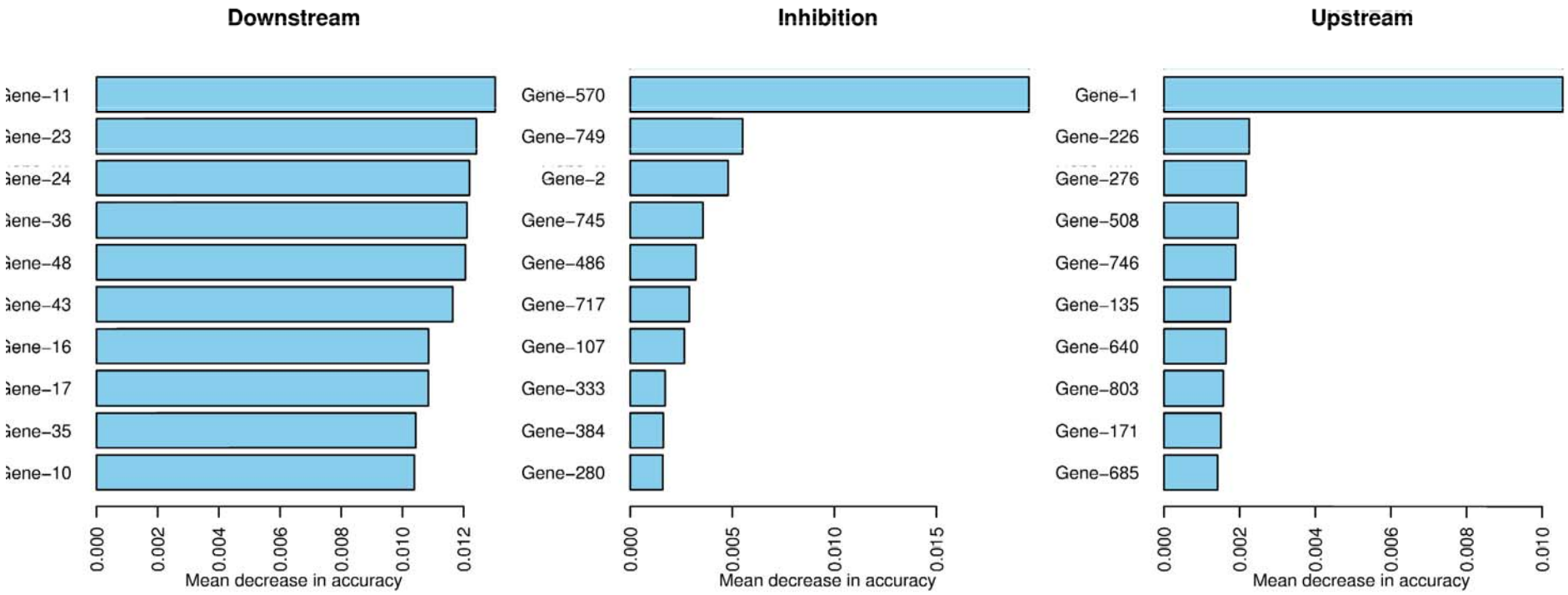
Downstream



Upstream



Random Forest



Conclusions

1. Statisticians should keep biology/hypotheses in mind when applying classification algorithms on Omics data

2. There are three main reasons why multiple markers outperform a single marker.
   - Downstream signalling
   - Inhibition/catalyzation
   - Different upstream causes

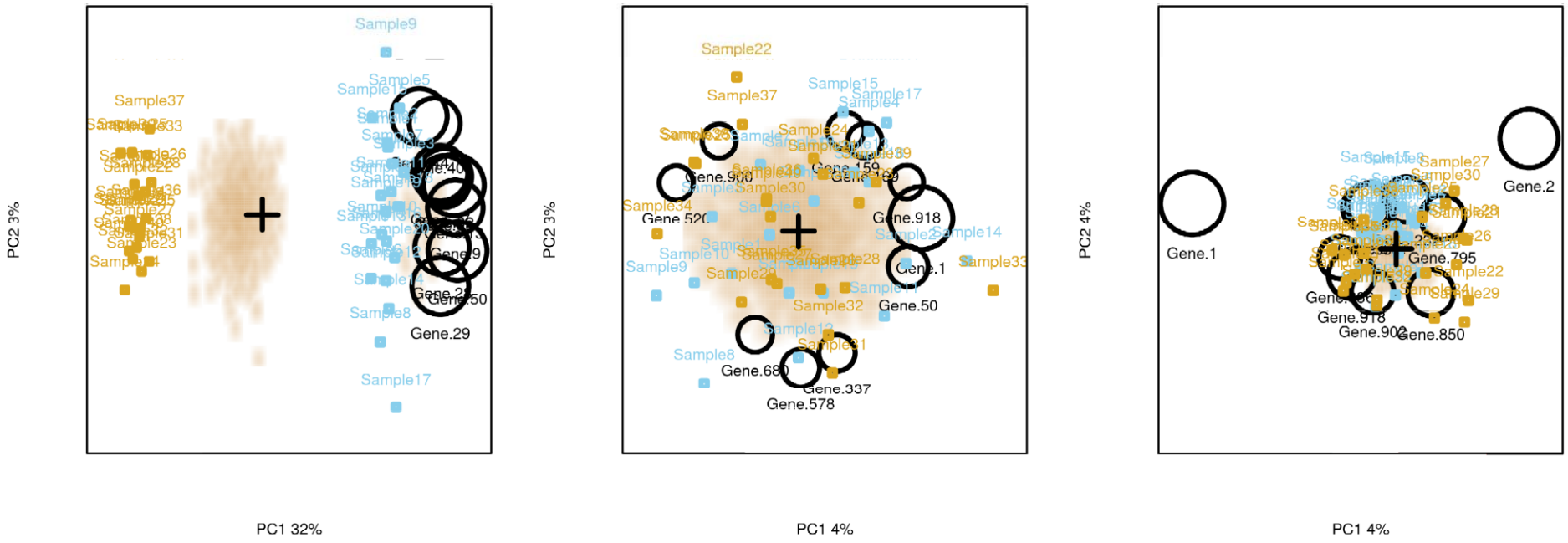
3. There are different statistical algorithms to adress each of these distinct hypotheses.

Thank you

Consequence of correlated features

- When going for ‘additive effects’, only one feature will be selected from a pool of correlated features.
  - Strong correlations between genes lead to non-unique solutions, impeding the biological interpretation of the obtained signature
- Going for ‘weighted average’ takes advantage of the correlational structure to make the signature more robust

Spectral maps (PCA)



When do multiple markers outperform single markers?

- Average of many correlated features
  - Concordance amongst a broader set of biomarkers in a qualification paradigm will increase confidence, leading to accepted and integrated translational biomarker signals
- Multiple markers each explain a different cause
  - A single marker is too simplistic and not complete